Analytical Modeling of Garbage Collection Algorithms in Hotness-aware Flash-based Solid State Drives

Yue Yang Jianwen Zhu

Electrical & Computer Engineering

University of Toronto

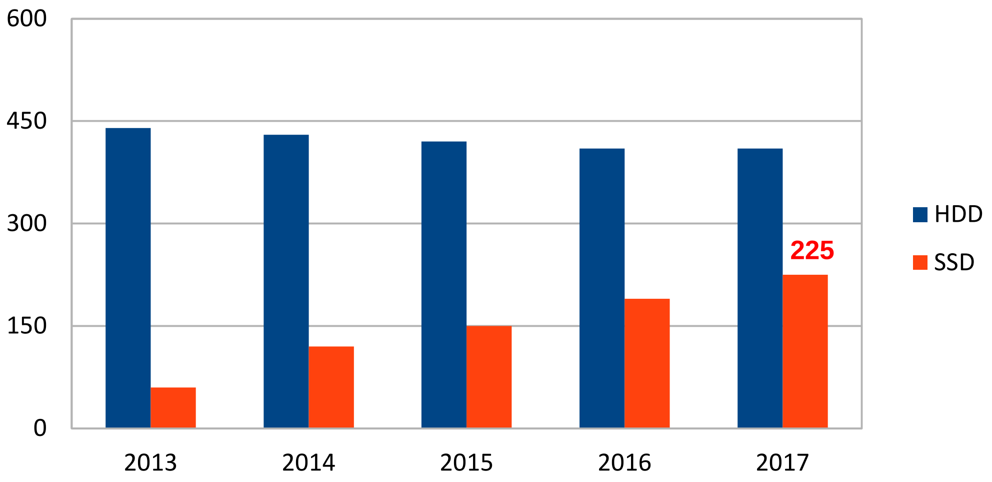
Agenda

- Introduction & motivation
- Analytical modeling
- Model validation
- Conclusion

Introduction

Solid State Drive Market Potential

Worldwide Shipment Forecast for SSDs and HDDs in PCs (Millions of units)



Source: IHS iSuppli Storage Market Tracker Report, May 2013

Technical Advantages

- Access latency ↓
- Bandwidth ↑
- Data safety ↑
- Power efficiency ↑
- Noise ↓

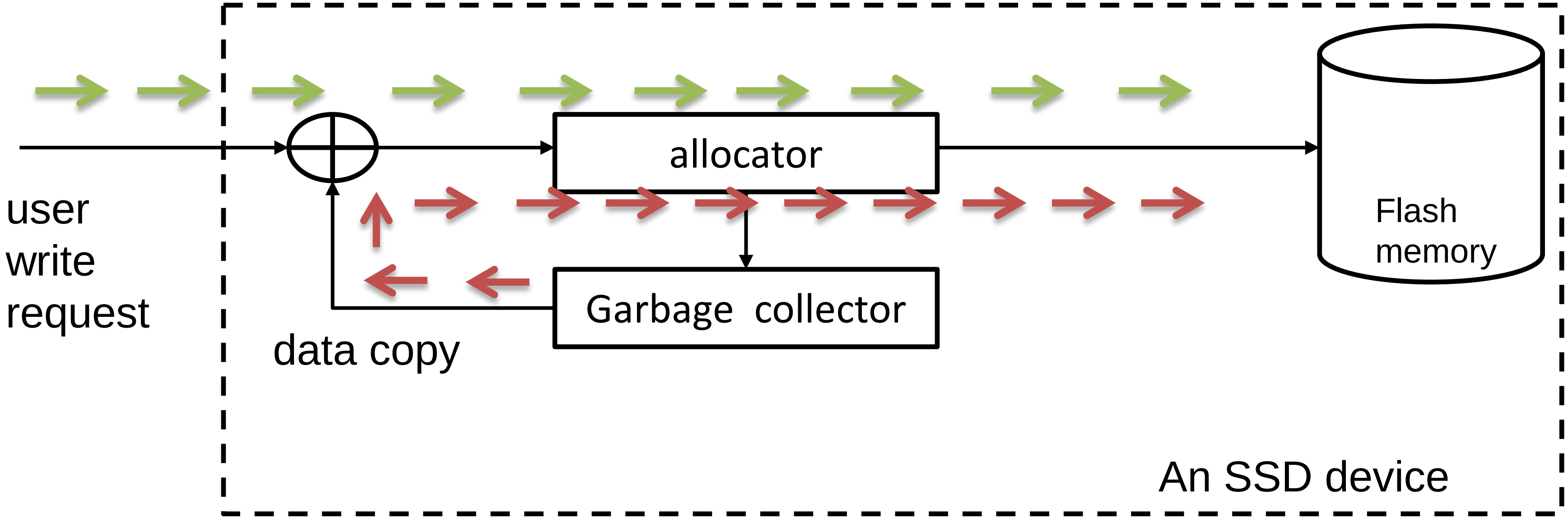
Flash Memory Limitation

- Endurance
  - limited budget of erase cycles (1K – 100K)
  - “erase-before-write” limitation

- *Question: How long will an SSD device last?*

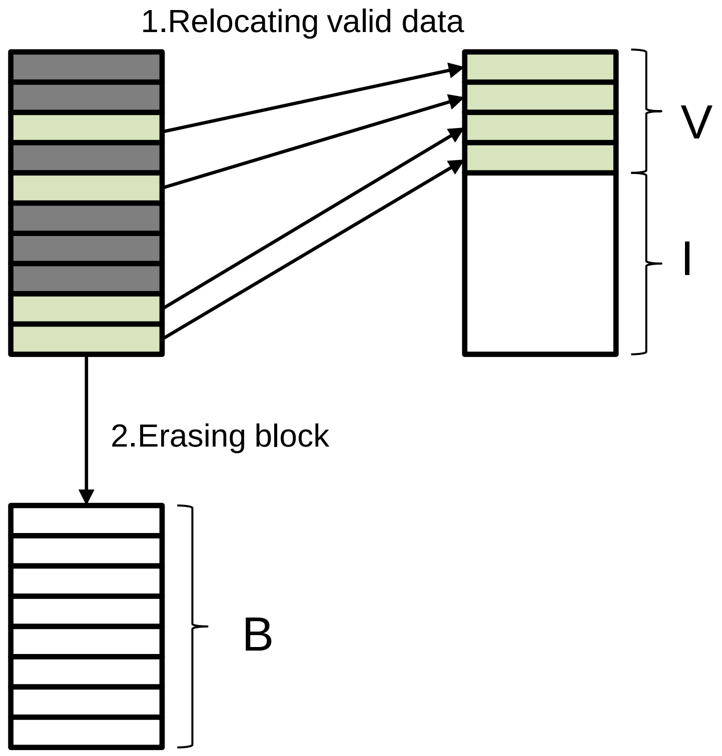
(how many user write requests can be serviced?)

Write Amplification



Garbage Collection

- Cleaning process
  - trigger condition
  - victim block selection
  - valid data migration (*source of write amplification*)
  - victim block erase



- Write amplification

$$A = \frac{B}{I} = \frac{B}{B - \overline{V}}$$

State-of-the-art

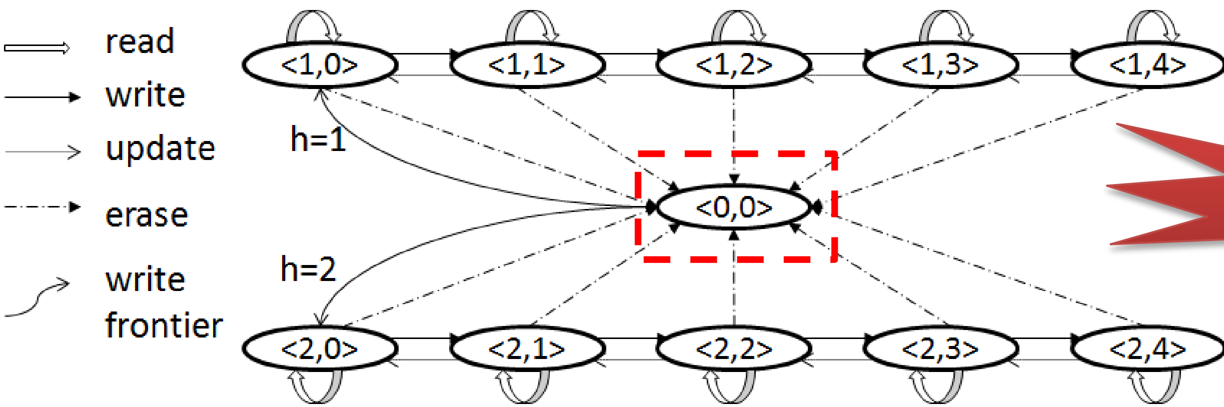
Framework	Workload model	Hotness separation	GC selection algorithm	Trace-driven validation
Bux (Perf.Eval'10)	Uniform	no	greedy	no
Houdt, (SIGMETRICS'13)	Uniform	no	d-Choice	no
Houdt, (Perf.Eval'13)	Hyper-exponential	no	d-Choice	no
Desnoyers, (SYSTOR'12)	Hyper-exponential	yes	greedy	yes
Li, (SIGMETRICS'13)	Poisson	no	d-Choice	yes
The proposed	general	yes	d-Choice	yes

Agenda

- Introduction & motivation
- Analytical modeling
- Model validation
- Conclusion

Life of An Erase Block

- Type of a <u>single</u> block *<h,v>*
  - h: the hotness tier that the block is allocated for
  - v: the number of valid pages in the block



state explosion

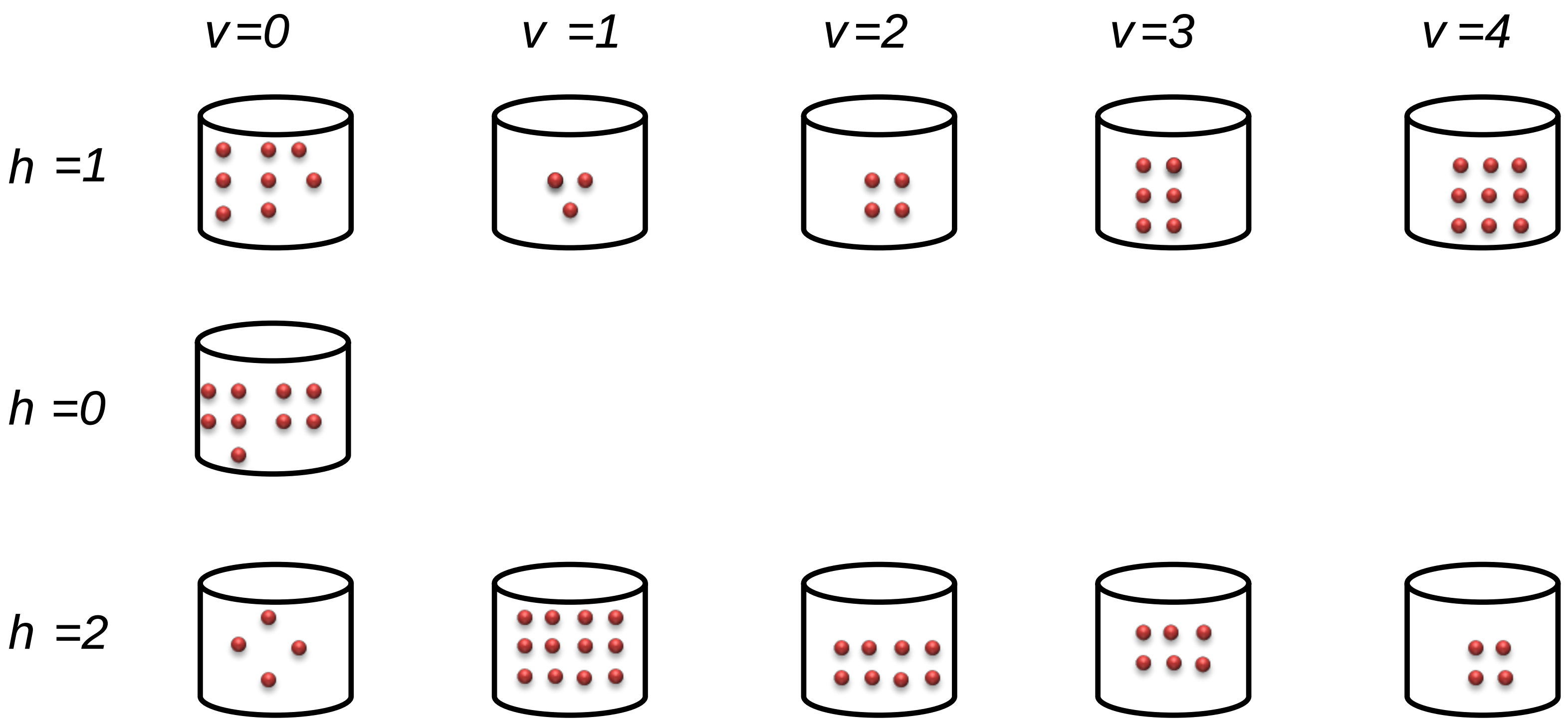
erased → Allocated (tier 1) → written → written → written → written

↓

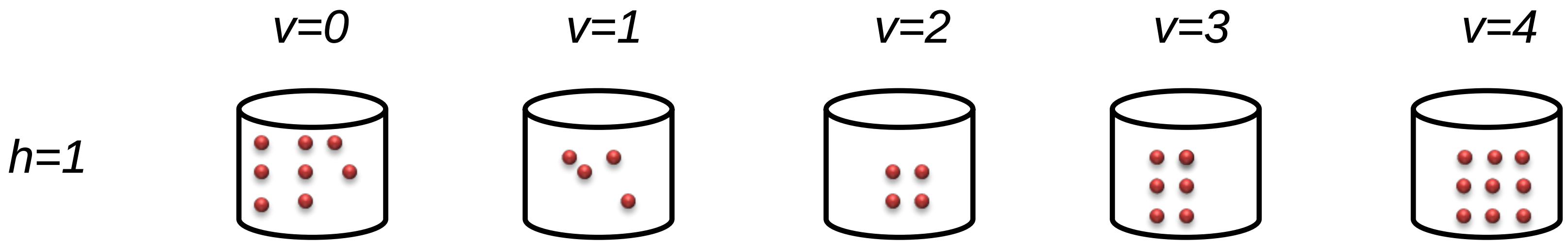
…… Allocated (tier 2) ← erased ← updated

System Dynamics

- State descriptor: occupancy measure vector $\vec{m}$
  - element : fraction of block type <*h,v*>
  - *Cardinality of* $\vec{m}$ *:* $|\mathcal{H}| \times |\mathcal{B}| + 1$



Event 1 – External Write Requests



P[a valid page in a <*h,v*> block is updated by an external write] =

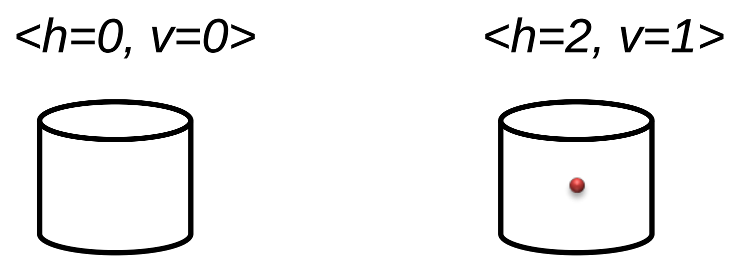
total number of valid physical pages in <h,v> blocks

$$u(h, v, \vec{m}) = r(h) \times \frac{m_v^h \times N \times (1-G) \times v}{B \times W \times f(h)}$$

Fraction of requests for tier h

total number of tier h logical pages

Event 2 – Block Erase



P[a <*h*,*v*> block is chosen by d-Choice as the victim] =

$$p(h,v,m) = \frac{m_v^h}{\sum_{h'=1}^{n} m_v^{h'}} \left[\left(\sum_{h'=1}^{n} \sum_{k=v}^{B} m_k^{h'} \right)^d - \left(\sum_{h'=1}^{n} \sum_{k=v+1}^{B} m_k^{h'} \right)^d \right]$$

The probability of selected block is of type <h,v>

The probability that all selected blocks have at least *v* valid pages

The probability that all selected blocks have at least *v+1* valid pages

A System of ODEs

For $0 \leq v \leq B$ and $1 \leq h \leq n$, let $g_v^h = \sum_{k=v}^{B} m_k^h$,

$$\Delta(g_v^h) = \underbrace{\sum_{k=0}^{v-1} p(h,k,\vec{m})}_{\text{increment rate of } g_v^h} - \underbrace{\left[B - \sum_{v=1}^{B}\left(\sum_{h'=1}^{n} g_v^{h'}\right)^{d}\right] \times u(h,v,\vec{m})}_{\text{decrement rate of } g_v^h} \qquad (1)$$

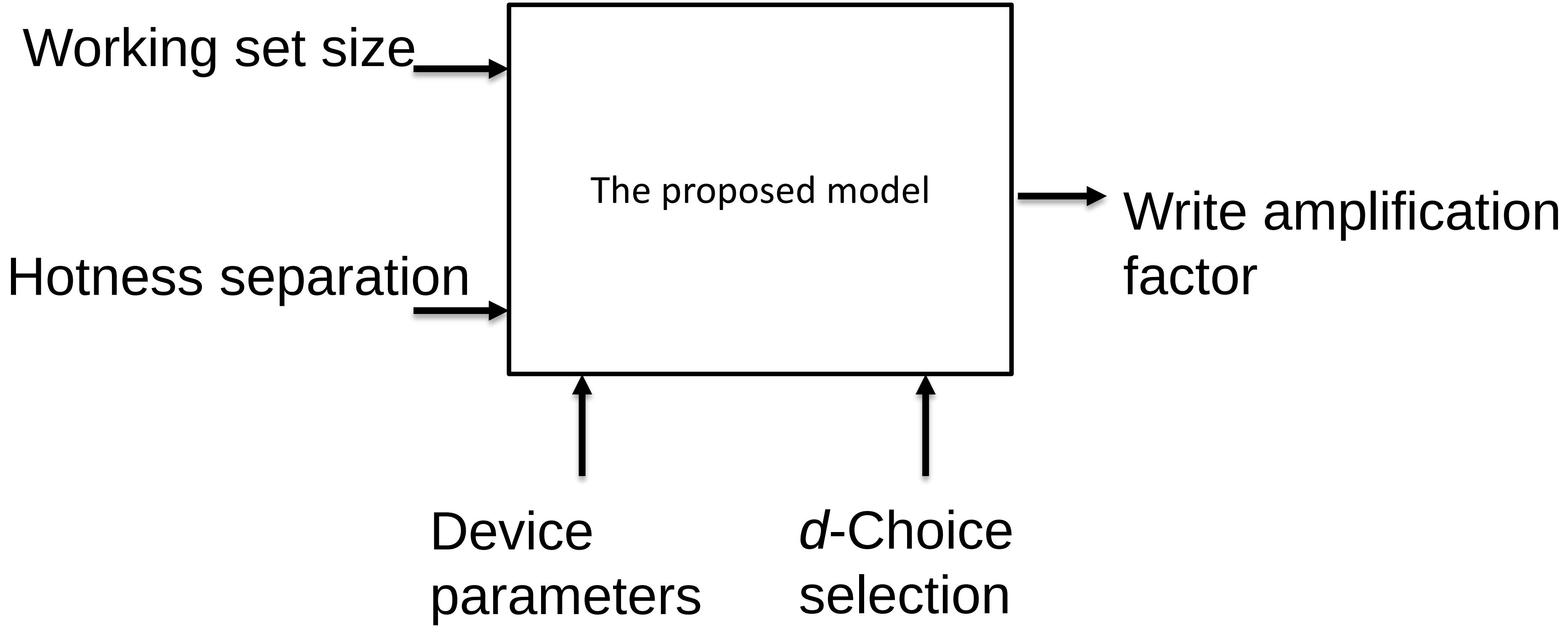
increment rate of g_v^h

decrement rate of g_v^h

<u>Mean field analysis & rescaling</u>

[1] Van Houdt, Benny. A Mean Field Model for a Class of Garbage Collection Algorithms in Flash-based Solid State Drives, sigmetric'13

Model Input / Output



Agenda

- Introduction & motivation
- Analytical modeling
- Model validation
- Conclusion

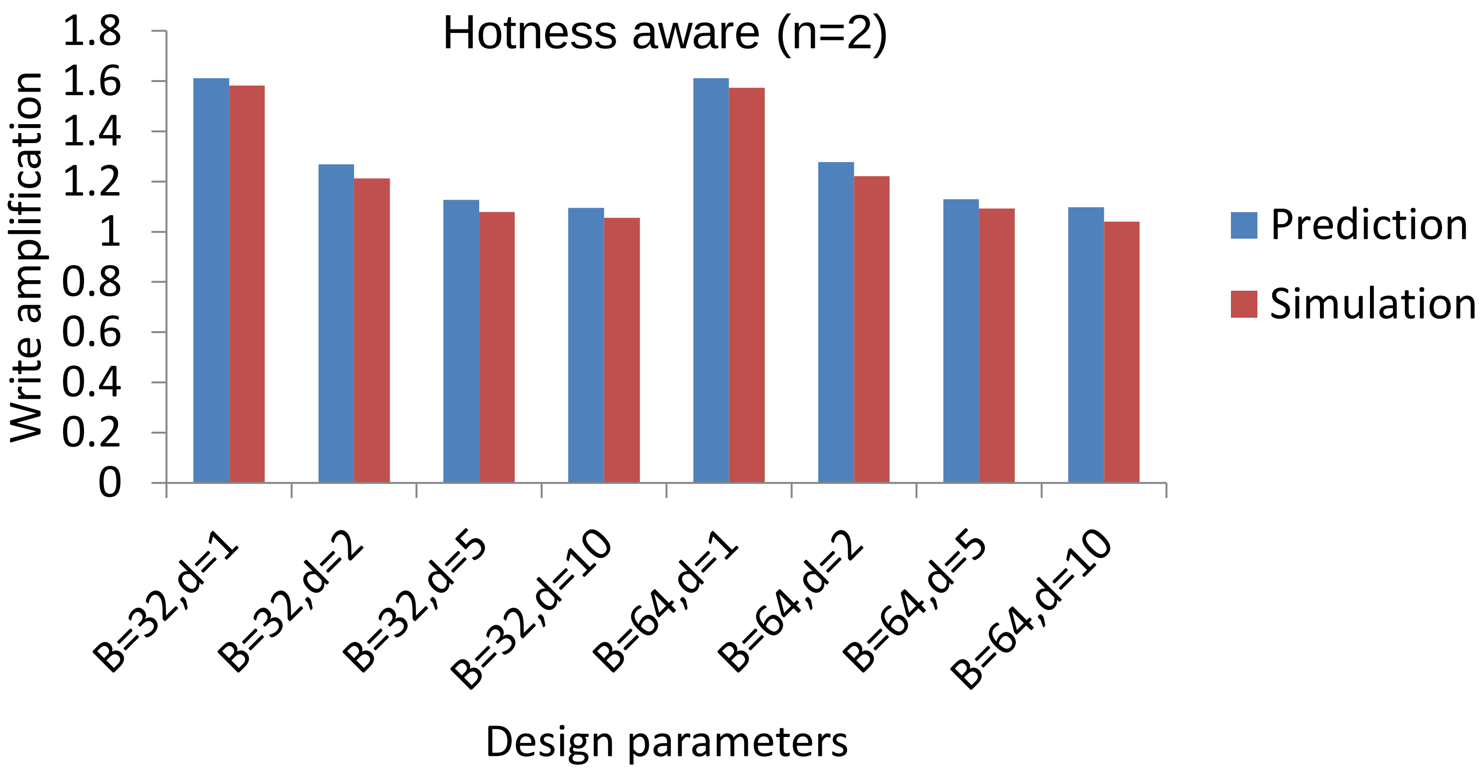
Simulation Setup

- The simulator
  - terabyte scale
  - highly configurable
  - trace-driven
- Run-time behavior
  - warm-up
  - statistics collection

Data Set

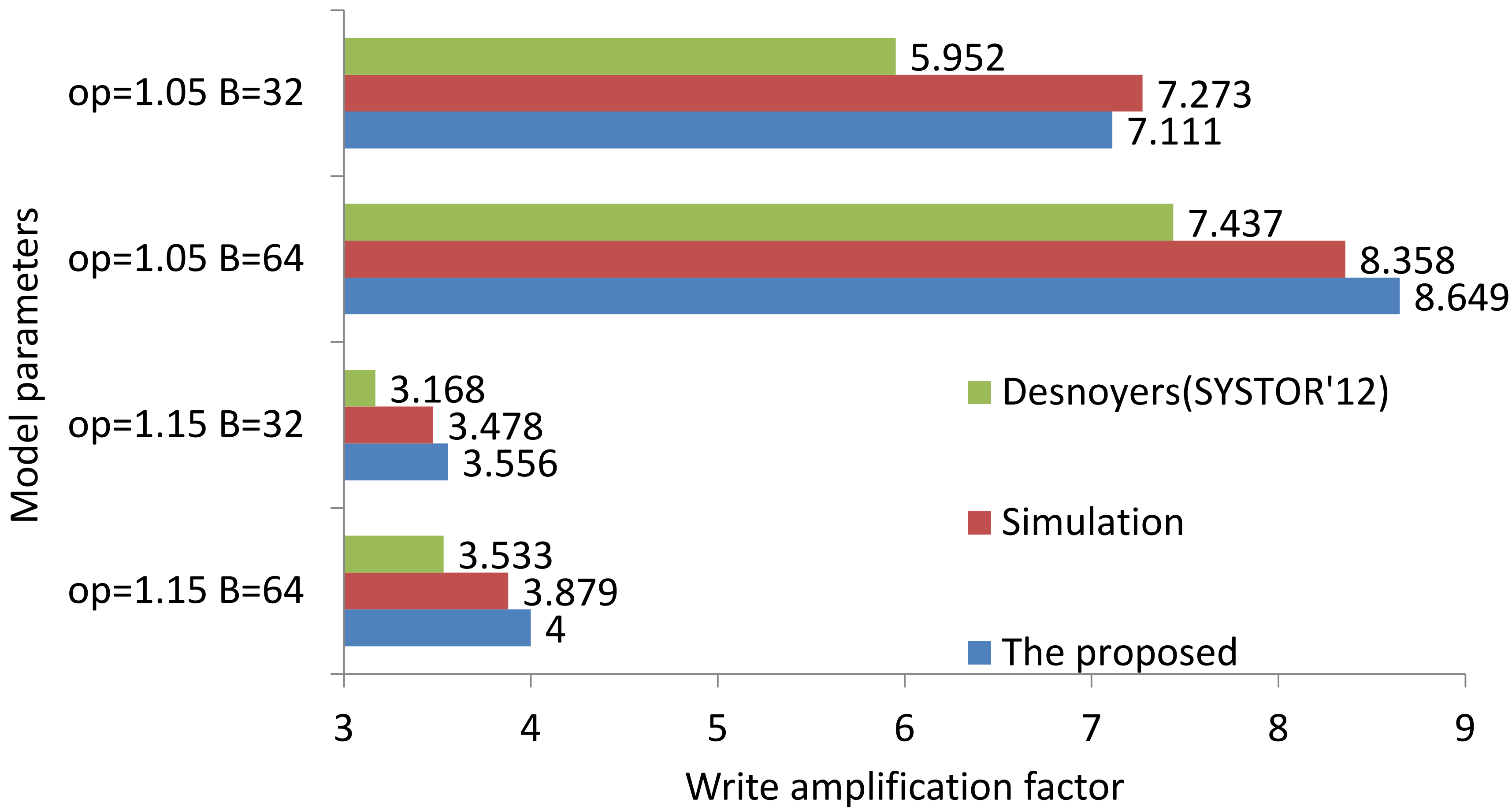
- FileBench synthetic traces
  - fileserver
  - OLTP
  - mail server
  - video server
  - web proxy
  - web server
- Real traces
  - OLTP application from a financial institution
  - Hardware monitor server in MS research, Cambridge

Prediction vs Simulation Result



Financial trace 2
-- Storage Performance Council. OLTP Application I/O. http://traces.cs.umass.edu/index.php/Storage/Storage, 2002.

Result Improvement



Write amplification prediction for greedy GC algorithm and hotness awareness.

Result Agreement

Hotness unaware write amplifications
Block size = *64*

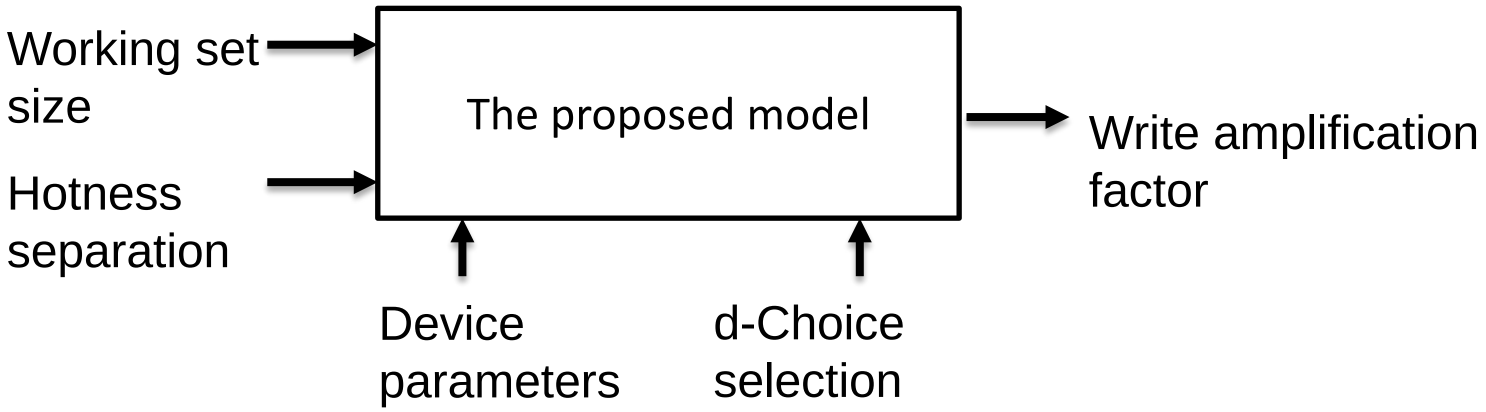
d	over-provisioning	The proposed	Houdt (SIGMETRICS'13)
2	1.07	9.63	9.64
4	1.07	7.72	7.72
8	1.07	7.00	7.00
2	1.16	4.96	4.96
4	1.16	4.08	4.07
8	1.16	3.73	3.74
2	1.26	3.37	3.37
4	1.26	2.80	2.80
8	1.26	2.59	2.59

Agenda

- Introduction & motivation
- Analytical modeling
- Model validation
- Conclusion

Conclusion

- An analytical model that can accurately predict the GC cleaning performance for
  - a general workload model
  - a wider class of selection algorithms
  - a write-frontier based hotness separation scheme



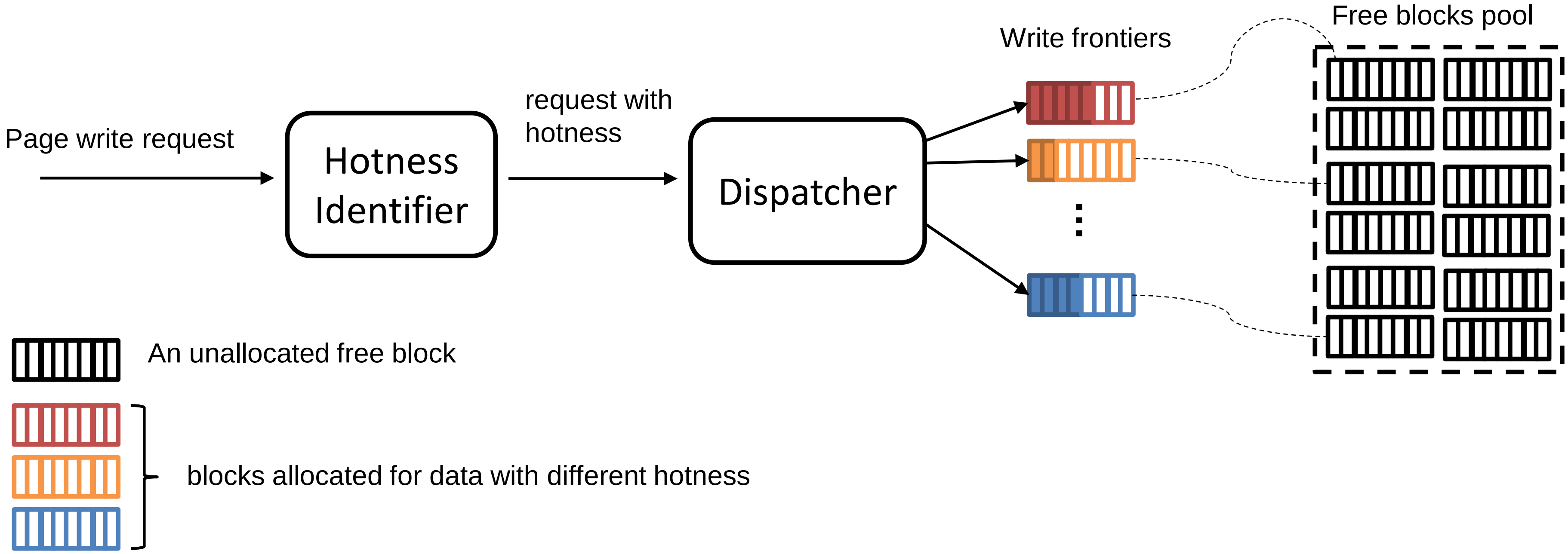
Thank you!

Contact :
Yue Yang yyang@eecg.toronto.edu
Jianwen Zhu jzhu@eecg.toronto.edu

Department of Electrical & Computer Engineering
University of Toronto

Allocator with Write Frontiers

- Hotness-based allocation
  - write frequency
  - one write frontier designated for each tier



Event – Internal Data Migration

- Valid data are always copied to a clean frontier
- Blocks state do not change